Inter-Rater Reliability of Historical Data Collected by Non-Medical Research Assistants and Physicians in Patients with Acute Abdominal Pain

Angela M. Mills, MD
Anthony J. Dean, MD
Frances S. Shofer, PhD
Judd E. Hollander, MD
Christine M. McCusker, RN
Michael K. Keutmann, BA
Esther H. Chen, MD

University of Pennsylvania, Department of Emergency Medicine

Supervising Section Editor: Craig L. Anderson, MPH, PhD
Submission history: Submitted April 15, 2008; Revision Received August 26, 2008; Accepted November 09, 2008.
Reprints available through open access at www.westjem.org

Objectives: In many academic emergency departments (ED), physicians are asked to record clinical data for research that may be time consuming and distracting from patient care. We hypothesized that non-medical research assistants (RAs) could obtain historical information from patients with acute abdominal pain as accurately as physicians.

Methods: Prospective comparative study conducted in an academic ED of 29 RAs to 32 resident physicians (RPs) to assess inter-rater reliability in obtaining historical information in abdominal pain patients. Historical features were independently recorded on standardized data forms by a RA and RP blinded to each others' answers. Discrepancies were resolved by a third person (RA) who asked the patient to state the correct answer on a third questionnaire, constituting the "criterion standard." Inter-rater reliability was assessed using kappa statistics (κ) and percent crude agreement (CrA).

Results: Sixty-five patients were enrolled (mean age 43). Of 43 historical variables assessed, the median agreement was moderate (κ 0.59 [Interquartile range 0.37-0.69]; CrA 85.9%) and varied across data categories: initial pain location (κ 0.61 [0.59-0.73]; CrA 87.7%), current pain location (κ 0.60 [0.47-0.67]; CrA 82.8%), past medical history (κ 0.60 [0.48-0.74]; CrA 93.8%), associated symptoms (κ 0.38 [0.37-0.74]; CrA 87.7%), and aggravating/alleviating factors (κ 0.09 [-0.01-0.21]; CrA 61.5%). When there was disagreement between the RP and the RA, the RA more often agreed with the criterion standard (64% [55-71%]) than the RP (36% [29-45%]).

Conclusion: Non-medical research assistants who focus on clinical research are often more accurate than physicians, who may be distracted by patient care responsibilities, at obtaining historical information from ED patients with abdominal pain.
[*West*JEM. 2009;10:30-36.]

INTRODUCTION

A busy emergency department (ED) is a challenging site for collecting data for prospective clinical trials. Frequently, treating physicians are asked to enroll eligible patients and complete structured data forms, a time-consuming process that can interfere with clinical responsibilities. Research assistants (RAs) without formal medical training [e.g., undergraduate and post-baccalaureate students] have been used to assist

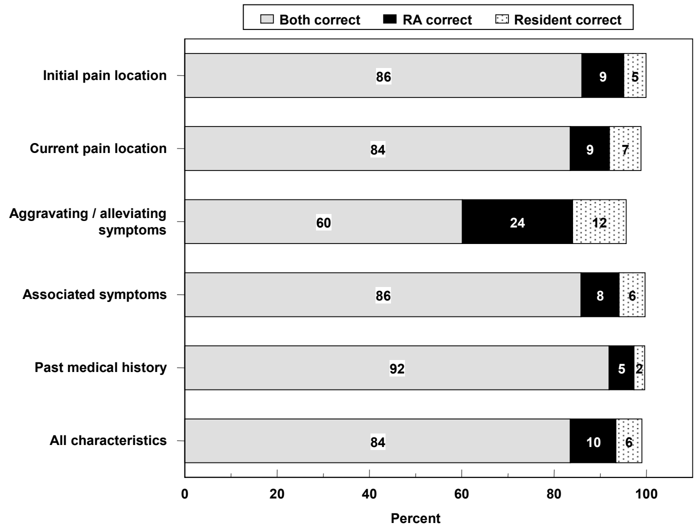

Figure 1. Accuracy of historical features by research assistants and physicians

in this process by identifying eligible patients, obtaining consent, documenting demographic information on standard data forms, and assisting with other data collection and management.[1-3]

Historical and physical examination features remain the basis for decision making about work-up and treatment of patients with acute abdominal pain; therefore, they are usually considered to be essential variables in research on this topic. Several studies have suggested that historical information obtained by medical providers may have significant inter-observer variability. In one such study, information recorded on standardized data sheets in a cohort of stroke patients revealed significant discrepancies in historical elements taken by six neurologists.[4] In a study of chest pain patients, the historical features documented by nurse practitioners were less typical of angina pectoris compared to those documented by physicians after interviewing the same patients.[5] These studies highlight the importance of assessing the reliability of the data- collection instrument as an integral part of the research project.

No study to date has examined the reliability of the non-medical RAs in obtaining historical information for research. We designed and piloted a survey instrument containing standard, simple historical questions about abdominal pain. We hypothesized that non-medical RAs can reliably use this questionnaire and be at least as accurate as resident physicians (RPs) in obtaining historical information from patients with acute abdominal pain.

METHODS

Study Design

We conducted a prospective comparative study to evaluate the reliability of the historical features obtained from ED patients with abdominal pain using a standard questionnaire administered by RAs compared to RPs. Our Institutional Committee on Research involving Human Subjects at the University of Pennsylvania approved the study. Informed consent was obtained from all subjects.

Study Setting and Population

This study was conducted at an urban university hospital ED with a annual census of approximately 55,000 visits. Adult patients with acute abdominal pain were enrolled from April 6 to 22, 2007. A survey instrument with questions about historical features was completed independently for each patient by a RA and a RP. RAs are undergraduate and post-baccalaureate students enrolled in the Academic Associate Program,[3, 6] a structured class at the University of Pennsylvania for which course credit is given. Students are responsible for attending research-related classes and working shifts in the ED during which they identify and enroll eligible patients for research projects, and in the current study, obtain historical information about patients with acute abdominal pain.

Study Protocol and Measurements

From 7 AM-midnight, seven days per week, the RAs identified and enrolled patients 18 years of age or older who presented with non-traumatic abdominal pain of less than 72 hours duration. Patients were excluded if they were pregnant, or if within the previous seven days they had sustained abdominal trauma or had an abdominal surgical procedure. A standardized questionnaire was completed independently by the RA and RP caring for the patient within 20 minutes of each other. The time of assessment was recorded on the data forms. Discrepancies between the two forms were resolved by a third person (RA) who was coached to specifically ask the patient: “we did not have a clear understanding of your answer to this question … [question repeated],” thus allowing patients to use either of their previous responses. This form was used as the “criterion standard.” Formal training sessions were provided to the RAs teaching them open-ended and neutral questioning techniques most likely to avoid influencing respondents.

Data Analysis

Descriptive data are presented as means ± standard deviation, frequencies, and percentages. Cohen’s kappa (κ) statistic and percent crude agreement (CrA), both with 95% confidence intervals (95% CIs), were used to measure inter-rater reliability. As described elsewhere, κ values range between 0 (chance agreement) and 1.00 (complete agreement); $\kappa<0.2$ represents poor agreement, 0.21-0.40 fair agreement, 0.41-0.60 moderate agreement, 0.61-0.80 good agreement, and 0.81-1.00 excellent agreement.[7] To summarize specific types of questions (e.g., past medical history) we present the median

Table 1. Kappa statistics and crude agreement for abdominal pain characteristics

Characteristics	Kappa	95% CI		Crude agreement	95% CI	
Initial pain location						
Pain start RUQ*	0.73	0.54	0.92	89.2%	79.1%	95.6%
Pain start LUQ*	0.61	0.40	0.82	83.1%	71.7%	91.2%
Pain start RLQ*	0.60	0.41	0.79	80.0%	69.1%	89.2%
Pain start LLQ*	0.57	0.37	0.77	78.5%	66.5%	87.7%
Pain start epigastrium	0.77	0.57	0.96	92.3%	83.0%	97.5%
Pain start both lower quadrants	0.79	0.64	0.95	90.8%	81.0%	96.5%
Pain start diffuse	0.66	0.39	0.94	92.3%	83.0%	97.5%
Pain start right flank	0.23	-0.04	0.50	80.0%	69.1%	89.2%
Pain start left flank	0.59	0.34	0.85	87.7%	77.2%	94.5%
Median and IQR*	**0.61 (0.59-0.73)**			**87.7% (80.0-90.8%)**		
Current pain location						
Pain now RUQ	0.673	0.482	0.864	85.9%	75.0%	93.4%
Pain now LUQ	0.602	0.401	0.802	81.3%	69.5%	69.5%
Pain now RLQ	0.594	0.398	0.790	79.7%	67.8%	88.7%
Pain now LLQ	0.466	0.248	0.683	73.4%	60.9%	83.7%
Pain now epigastrium	0.656	0.442	0.870	87.5%	76.9%	94.5%
Pain now both lower quadrants	0.692	0.508	0.877	85.9%	75.0%	93.4%
Pain now diffuse	0.744	0.508	0.979	93.8%	84.8%	98.3%
Pain now right flank	0.455	0.184	0.726	82.8%	71.3%	91.1%
Pain now left flank	0.301	0.009	0.592	81.3%	69.5%	69.5%
Median and IQR	**0.60 (0.47-0.67)**			**61.5% (53.1-67.0%)**		
Aggravating/alleviating symptoms						
Pain ever gone	0.237	-0.010	0.480	68.3%	55.3%	79.4%
Eating aggravating	0.130	-0.050	0.310	50.8%	38.1%	63.4%
Urinating aggravating	0.040	-0.160	0.240	76.9%	64.8%	86.5%
Coughing aggravating	0.310	0.090	0.520	63.1%	50.2%	74.7%
Antacid alleviating	-0.050	-0.270	0.160	41.5%	29.4%	54.4%
Eating alleviating	-0.020	-0.200	0.160	60.0%	47.1%	72.0%
Median and IQR	**0.09(-0.01-0.21)**			**61.5% (53.1-67.0%)**		
Associated symptoms						
Vomiting	0.740	0.570	0.900	87.7%	77.2%	94.5%
Diarrhea	0.780	0.600	0.960	92.3%	83.0%	97.5%
Dysuria	NC*			96.9%	89.3%	99.6%
Pass gas	0.160	-0.090	0.400	60.0%	47.1%	72.0%
Fever	0.380	0.150	0.610	73.8%	61.5%	84.0%
Vaginal discharge	NC*			96.0%	86.3%	99.5%
Vaginal bleeding	0.370	-0.190	0.930	94.0%	83.5%	98.8%
Median and IQR	**0.38 (0.37-0.74)**			**87.7% (73.8-93.2%)**		

Table 1. Kappa statistics and crude agreement for abdominal pain characteristics

Characteristics	Kappa	95% CI		Crude agreement	95% CI	
Past Medical History						
HX* Abdominal surgery	0.690	0.520	0.870	84.6%	73.5%	92.4%
HX Gallstones	0.580	0.260	0.900	92.3%	83.0%	97.5%
HX Liver Disease	0.500	0.130	0.880	92.3%	83.0%	97.5%
HX Pancreatitis	1.000	1.000	1.000	100.0%	94.5%	100.0%
HX Inflammatory bowel disease	1.000	1.000	1.000	100.0%	94.5%	100.0%
HX Irritable bowel syndrome	0.420	0.020	0.820	92.3%	83.0%	97.5%
HX Diverticulitis	0.550	0.090	1.000	95.4%	87.1%	99.0%
HX GERD*	0.210	0.000	0.420	72.3%	59.8%	82.7%
HX Kidney stones	0.610	0.390	0.830	86.2%	75.3%	93.5%
HX Cancer	0.870	0.700	1.000	96.9%	89.3%	99.6%
HX Diabetes	0.700	0.390	1.000	95.4%	87.1%	99.0%
HX CAD*	0.380	-0.180	0.930	95.4%	87.1%	99.0%
Median and IQR	**0.60 (0.48-0.74)**			**93.8% (85.9-95.8%)**		
Overall Median	**0.59 (0.37-0.69)**			**85.9% (77.7-92.3%)**		

RUQ, right upper quadrant; *LUQ,* left upper quadrant; *RLQ,* right lower quadrant; *LLQ,* left lower quadrant; *IQR,* interquartile range; *NC,* not calculable; *HX,* history; *GERD,* gastroesophageal reflux disease; *CAD,* coronary artery disease.

kappa values with interquartile ranges (IQRs). Data were analyzed using SAS statistical software (Version 9.1, SAS Institute, Cary, NC) and StatXact (Version 6.1, Cytel Software Corporation, Cambridge, MA).

RESULTS

Sixty-five patients with acute abdominal pain were surveyed by 29 RAs and 32 RPs. The median age of the abdominal pain patients was 43 years; 77% were female and 54% black. There were 49 variables, of which 43 were dichotomized responses. The remaining six historical variables were related to times (e.g. when was the last time you vomited), which proved highly variable and not easily dichotomized. These were excluded. Therefore, there were 2754 comparisons (some variables had fewer comparisons and some were restricted by gender), of which there were 458 discrepancies between RP and RA (17%).

Inter-rater reliability measures for all historical variables are listed in Table 1. Overall, the median agreement was moderate (κ 0.59 [IQR 0.37-0.69]; CrA 85.9%) but varied across data categories: initial pain location (κ 0.61 [IQR 0.59-0.73]; CrA 87.7%), current pain location (κ 0.60 [IQR 0.47-0.67]; CrA 82.8%), past medical history (κ 0.60 [IQR 0.48-0.74]; CrA 93.8%), associated symptoms (κ 0.38 [IQR 0.37-0.74]; CrA 87.7%), and aggravating/alleviating factors (κ 0.09 [IQR -0.01-0.21]; CrA 61.5%).

Overall, crude agreement for both groups was above 80% in all but one of the five general categories (Figure 1). Of the 458 discordant results between the RP and RA, criterion standard was available for 429 (94%). Of these disagreements, the RA more often agreed with the criterion standard (N=274, 64% [55%-71%] compared with the RP (N=155, 36% [29-45%]. (See Table 2.)

DISCUSSION

This study explores the inter-rater reliability of historical features obtained by RAs and RPs using a standard questionnaire in the evaluation of abdominal pain. We found an overall moderate agreement between RAs and RPs for 43 historical variables. There was good agreement for initial pain location and moderate agreement for current pain location and past medical history. For associated symptoms, there was fair agreement using the kappa statistic with a crude agreement of 88%. The poorest agreement was found for aggravating and alleviating factors in which information obtained by both groups of investigators was correct only 62% of the time. The mathematical properties of the κ statistic determine that low rates of discrepancy in infrequent clinical findings will result in lower κ scores than the same rate in common ones. This may have resulted in the wide range of alleviating and aggravating factors, any one of which is encountered relatively infrequently, appearing to result in lower κ scores.

Our results are consistent with prior studies of inter-rater reliability of physicians obtaining historical features,

Table 2. Accuracy amongst discordant pairs compared to criterion standard

Characteristics	Number discordant pairs	%RA correct	%RP correct
Initial pain location			
Pain start RUQ*	7	71.4%	28.6%
Pain start LUQ*	11	63.6%	36.4%
Pain start RLQ*	12	50.0%	50.0%
Pain start LLQ*	14	57.1%	42.9%
Pain start epigastrium	5	100.0%	0.0%
Pain start both lower quadrants	6	50.0%	50.0%
Pain start diffuse	5	60.0%	40.0%
Pain start right flank	13	76.9%	23.1%
Pain start left flank	8	75.0%	25.0%
Median and IQR*		**63.6% (57.1-75.0%)**	**36.4% (25.0-42.9%)**
Current pain location			
Pain now RUQ	8	37.5%	62.5%
Pain now LUQ	11	45.5%	54.5%
Pain now RLQ	13	53.8%	46.2%
Pain now LLQ	16	56.3%	43.8%
Pain now epigastrium	7	42.9%	57.1%
Pain now both lower quadrants	8	75.0%	25.0%
Pain now diffuse	3	66.7%	33.3%
Pain now right flank	10	60.0%	40.0%
Pain now left flank	12	66.7%	33.3%
Median and IQR		**56.3% (45.5-66.7%)**	**43.8% (33.3-54.5%)**
Aggravating/alleviating symptoms			
Pain ever gone	19	63.2%	36.8%
Eating aggravating	27	74.1%	25.9%
Urinating aggravating	14	64.3%	35.7%
Coughing aggravating	21	71.4%	28.6%
Antacid alleviating	36	63.9%	36.1%
Eating alleviating	21	66.7%	33.3%
Median and IQR		**65.6% (64.0-70.2%)**	**34.5% (29.8-36.0%)**
Associated symptoms			
Vomiting	7	71.4%	28.6%
Diarrhea	5	60.0%	40.0%
Dysuria	2	100.0%	0.0%
Pass gas	26	65.4%	34.6%
Fever	17	41.2%	58.8%
Vaginal discharge	2	50.0%	50.0%
Vaginal bleeding	3	66.7%	33.3%
Median and IQR		**65.4% (55.0-69.0%)**	**34.6% (31.0-45.0%)**

Table 2. Accuracy amongst discordant pairs compared to criterion standard

Characteristics	Number discordant pairs	%RA correct	%RP correct
Past Medical History			
HX* Abdominal surgery	9	55.6%	44.4%
HX Gallstones	5	60.0%	40.0%
HX Liver Disease	5	100.0%	0.0%
HX Pancreatitis	0	no discordant pairs	no discordant pairs
HX Inflammatory bowel disease	0	no discordant pairs	no discordant pairs
HX Irritable bowel syndrome	4	100.0%	0.0%
HX Diverticulitis	3	0.0%	100.0%
HX GERD*	17	64.7%	35.3%
HX Kidney stones	9	77.8%	22.2%
HX Cancer	2	50.0%	50.0%
HX Diabetes	3	100.0%	0.0%
HX CAD*	3	100.0%	0.0%
Median and IQR		**62.4% (54.2-83.3%)**	**37.6% (16.7-45.8%)**
Overall Median		**63.9% (54.7-71.4%)**	**36.1% (28.6-45.3%)**

RUQ, right upper quadrant; *LUQ,* left upper quadrant; *RLQ,* right lower quadrant; *LLQ,* left lower quadrant; *IQR,* interquartile range; *NC,* not calculable; *HX,* history; *GERD,* gastroesophageal reflux disease; *CAD,* coronary artery disease.

showing fair to excellent agreement (κ range 0.27-0.89) in hospitalized chest pain patients,[8] fair to good agreement (κ range 0.37-0.69) in suspected stroke patients,[9] and good agreement (κ range 0.58-0.71) in patients with suspected osteoarthritis.[10] The current study also supports the findings of reports in which non-physician army medical practitioners demonstrated good overall agreement compared to physicians in the assessment of upper respiratory infection.[11, 12] Specific to abdominal pain, our results were also consistent with those of a recent study comparing pediatric emergency physicians with surgeons in the evaluation of appendicitis in children showing fair to excellent agreement (κ range 0.33-0.82) for historical questions.[13]

Accurate data collection is an essential component of high quality clinical research. Prospectively collected data is generally considered to be of higher quality than data collected retrospectively or through chart abstraction. In many prospective studies conducted in the ED, the treating physician is asked to record subjects' clinical data. This process may be cumbersome and time consuming. It may also be distracting or interfere with the physicians' other responsibilities or create a fundamental conflict between the physician's role as care provider and as researcher. To date, this is the first study to compare the ability of RAs with no formal medical training to RPs in obtaining historical information for research purposes. If, as the current study suggests, non-medical research assistants can obtain historical information about ED patients' acute abdominal pain that is as accurate or more accurate than that obtained by the treating physician, the burden of data collection may be lifted from the treating physician, allowing it to be obtained and recorded in a less hurried and more meticulous manner. This may result in higher quality medical research on this topic in the ED setting.

LIMITATIONS

As this study was conducted in a single institution with an established Academic Associate Program, our results may not be generalizable to other practice settings. Under-enrollment of patients evaluated in the overnight hours, the most acutely ill patients, and patients who did not consent to participate in the study may have caused some selection bias. The authors do not know of any "gold standard" available to be certain that patient responses to historical items are accurate. As such, this study design was our best attempt to study accuracy and inter-rater reliability in obtaining historical data for patients with abdominal pain. It is possible that patients may have been prompted into providing answers that were consistent with one of their prior responses when being interviewed by the third person for the "criterion standard" form. It is also possible that the third-person interviewer might have had a tendency to "coach" respondents to resolve discrepancies in a way that supported the data obtained by the first RA. Neither RAs nor RPs were blinded to the purpose of the study, which may have biased our results.

CONCLUSION

Non-medical research assistants focused on clinical research are often more accurate than physicians, who may be distracted by patient care responsibilities, at obtaining data for clinical research. They can reliably use a standardized data collection sheet to obtain historical information from patients who present to the ED with acute abdominal pain.

Address for Correspondence: Angela M. Mills, MD. Department of Emergency Medicine, Ground Floor, Ravdin Building, Hospital of the University of Pennsylvania, 3400 Spruce Street, Philadelphia, PA 19104-4283. Email: millsa@uphs.upenn.edu.

Conflicts of Interest: By the *West*JEM article submission agreement, all authors are required to disclose all affiliations, funding sources, and financial or management relationships that could be perceived as potential sources of bias. The authors disclosed none.

REFERENCES

1. Bradley K, Osborn HH, Tang M. College research associates: a program to increase emergency medicine clinical research productivity. *Annals of Emergency Medicine*. 1996; 28:328-333.
2. Cobaugh DJ, Spillane LL, Schneider SM. Research subject enroller program: a key to successful emergency medicine research. *Acad Emerg Med*. 1997; 4:231-233.
3. Hollander JE, Valentine SM, Brogan GX, Jr. Academic associate program: integrating clinical emergency medicine research with undergraduate education. *Acad Emerg Med*. 1997; 4:225-230.
4. Shinar D, Gross CR, Mohr JP, et al. Interobserver variability in the assessment of neurologic history and examination in the Stroke Data Bank. *Archives of Neurology*. 1985; 42:557-565.
5. Hickam DH, Sox HC, Jr., Sox CH. Systematic bias in recording the history in patients with chest pain. *Journal of Chronic Diseases*. 1985; 38:91-100.
6. Hollander JE, Singer AJ. An innovative strategy for conducting clinical research: the academic associate program. *Acad Emerg Med*. 2002; 9:134-137.
7. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ*. 1992; 304:1491-1494.
8. James TL, Feldman J, Mehta SD. Physician variability in history taking when evaluating patients presenting with chest pain in the emergency department. *Acad Emerg Med*. 2006; 13:147-152.
9. Hand PJ, Haisma JA, Kwan J, et al. Interobserver agreement for the bedside clinical assessment of suspected stroke. *Stroke*. 2006; 37:776-780.
10. Peat G, Wood L, Wilkie R, et al. How reliable is structured clinical history-taking in older adults with knee problems? Inter- and intraobserver variability of the KNE-SCI. J*ournal of Clinical Epidemiology*. 2003; 56:1030-1037.
11. Wilson FP, Wilson LO, Wheeler MF, et al. Algorithm-directed care by nonphysician practitioners in a pediatric population: Part I. Adherence to algorithm logic and reproducibility of nonphysician practitioner data-gathering behavior. *Medical Care*. 1983; 21:127-137.
12. Wood RW, Diehr P, Wolcott BW, et al. Reproducibility of clinical data and decisions in the management of upper respiratory illnesses: a comparison of physicians and non-physician providers. *Medical Care*. 1979; 17:767-779.
13. Kharbanda AB, Fishman SJ, Bachur RG. Comparison of pediatric emergency physicians' and surgeons' evaluation and diagnosis of appendicitis. *Acad Emerg Med*. 2008; 15:119-125.